Journal of the American Statistical Association

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/uasa20

Discussion of "A Tuning-Free Robust and Efficient Approach to High-Dimensional Regression"

Xiudi Li & Ali Shojaie

To cite this article: Xiudi Li & Ali Shojaie (2020) Discussion of "A Tuning-Free Robust and Efficient Approach to High-Dimensional Regression", Journal of the American Statistical Association, 115:532, 1717-1719, DOI: 10.1080/01621459.2020.1837139

To link to this article: https://doi.org/10.1080/01621459.2020.1837139

JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2020, VOL. 115, NO. 532, 1717–1719: Theory and Methods Discussion
https://doi.org/10.1080/01621459.2020.1837139

Discussion of "A Tuning-Free Robust and Efficient Approach to High-Dimensional Regression"

Xiudi Li and Ali Shojaie

Department of Biostatistics, University of Washington, Seattle, WA, USA

We commend the authors for their valuable contribution to the field of high-dimensional regression.

Despite considerable progress over the past two decades, a critical gap remains between the theory and applications of penalized estimation procedures for high-dimensional regression. On the one hand, the vast majority of existing theoretical developments have focused on (sub)Gaussian errors. However, outliers and heavy-tailed distributions naturally arise in many applications. On the other hand, despite recent theoretical developments (Sun and Zhang 2012; Lederer and Muller 2015; Sabourin, Valdar, and Nobel 2015, etc), tuning parameter selection remains challenging in practice. Although cross-validation is widely used and leads to great predictive performance, it may not be optimal for the purpose of variable selection (see, e.g., Meinshausen and Bühlmann 2006) or statistical inference.

The paper by Wang and co-authors bridges this critical gap between theory and practice by proposing a new rank-based procedure for high-dimensional regression, which requires minimal tuning and is robust to heavy-tailed error distributions. We believe this promising approach provides applied statisticians and practitioners with a robust and powerful alternative for high-dimensional regression and opens the door to future methodological and theoretical developments. To highlight potential extensions, in this discussion, we explore applications of the proposed approach in other contexts, in particular, graphical modeling (Maathuis et al. 2018). We also investigate the possibilities to develop valid statistical inference for the proposed estimators, which would be especially important for scientific applications.

1. Applications in Graphical Modeling

The proposed rank-based lasso procedure provides a robust and tuning-free estimator that could be used in many other applications involving high-dimensional regression. Here, we explore structure learning in high-dimensional graphical models (Drton and Maathuis 2017) as one such example.

Despite recent developments in graphical models for non-Gaussian observations (Voorman, Shojaie, and Witten 2014; Chen, Witten, and Shojaie 2015; Lin, Drton, and Shojaie 2016; Fan et al. 2017; Yu, Drton, and Shojaie 2019, etc.), the two most-widely used approaches for learning the structure of graphical models are *neighborhood selection*, for example, by doing node-wise regression with lasso (Meinshausen and Bühlmann 2006); and penalized maximum likelihood estimation, such as *graphical lasso* (Yuan and Lin 2007; Friedman, Hastie, and Tibshirani 2008). Both of these methods are designed primarily for Gaussian data. Graphical modeling based on copulas and rank correlation (Liu, Lafferty, and Wasserman 2009; Liu et al. 2012; Xue et al. 2012) provides more flexible alternatives in terms of the underlying distribution, but still require multivariate normality after monotone transformations. In addition, tuning parameter selection in the graphical model setting can be even more challenging. Combining node-wise regression with the proposed rank-based lasso offers an appealing alternative that overcome these difficulties.

Suppose we want to infer the conditional independence relationships among components of a random vector $X \in \mathbb{R}^p$. For $j = 1, \ldots, p$, we regress X_j against X_{-j} using the proposed rank-based lasso or rank-based SCAD methods (Wang et al. 2020), and let $\hat{\beta}_{jk}$ be the regression coefficient of X_k in this regression, for $k \neq j$. Let $\hat{E}$ be the estimated edge set. Then $(j, k) \in \hat{E}$ if either $\hat{\beta}_{jk}$ or $\hat{\beta}_{kj}$ is nonzero. Alternatively, one may choose to include an edge only when both $\hat{\beta}_{jk}$ and $\hat{\beta}_{kj}$ are nonzero.

We next present some preliminary simulation results to investigate the advantages of the proposed approach over existing procedures. Here, the data are generated from a nonparanormal distribution (Liu, Lafferty, and Wasserman 2009) with $p = 100$. The undirected conditional independence graph is estimated using the graphical lasso (glasso) (Yuan and Lin 2007; Friedman, Hastie, and Tibshirani 2008), the nonparanormal SKEPTIC (npn) (Liu et al. 2012; Xue et al. 2012), nodewise rank lasso, and nodewise rank SCAD. The tuning parameter is chosen by BIC for glasso and npn. While rank lasso is tuning-free, the second stage in rank SCAD requires some tuning and we use the hBIC as proposed by the authors (Wang et al. 2020).

Table 1 shows the average true positive rate (TPR) and false positive rate (FPR) of the above procedures over 100 replications, for $n = 100$ and 200. We observe that for glasso and npn, tuning parameter selection can be difficult and BIC often gives an empty graph. In contrast, rank lasso and rank SCAD require less tuning, and perform well even in this case where the linear model is misspecified. These preliminary results highlight

CONTACT Ali Shojaie ashojaie@uw.edu Department of Biostatistics, University of Washington, Seattle, WA 98195.

Table 1. True positive and false positive rate of graphical lasso, nonparanormal SKEPTIC, nodewise rank lasso and nodewise rank SCAD over 100 replications.

		glasso	npn	Rank lasso	Rank SCAD
$n = 100$	TPR	0 (0)	0 (0)	0.24 (0.03)	0.64 (0.04)
	FPR	0 (0)	0 (0)	0.00 (0.00)	0.14 (0.02)
$n = 200$	TPR	0 (0)	0.10 (0.27)	0.59 (0.04)	0.86 (0.03)
	FPR	0 (0)	0.00 (0.01)	0.00 (0.00)	0.10 (0.01)

NOTE: Data were generated from a nonparanormal distribution with $p = 100$.

the potential advantage of the proposed rank-based estimator in graphical modeling applications and other unsupervised learning settings, where tuning parameter selection is challenging and outliers or heavy-tailed distributions may adversely impact the performance of existing procedures.

2. Statistical Inference

Statistical inference is important for adoption of the proposed rank-based lasso in scientific applications, as practitioners often seek to quantify the uncertainty of estimates. The proposed rank lasso has the advantage of being tuning-free and robust to heavy-tailed error distributions, which would be desirable for inference. Here, we consider several possibilities for developing valid inference procedures that inherit these advantages, for the rank-based lasso.

As noted by the authors, minimizing the proposed loss is equivalent to minimizing the Jaeckel's dispersion function with Wilcoxon scores

$$L(\beta) = \sqrt{12}\sum_{i=1}^{n}\left[\frac{R(Y_i - x_i^T\beta)}{n+1} - \frac{1}{2}\right](Y_i - x_i^T\beta), \quad (1)$$

where $R(\cdot)$ denotes the rank. We work with this version of the loss function in this section.

2.1. Inference based on refitting

Theorem 2 shows that when the number of active variables is fixed, with high probability, the subvector of the second-stage estimator (rank SCAD) corresponding to the selected variables equals to the oracle estimator. Given this result, the authors then note that this subvector has an asymptotically normal distribution. One can thus consider a Wald-type inference procedure based on rank SCAD directly. However, in our simulation experiments, we observed that even with moderately large samples, the distribution of rank SCAD can have a small bias and is often skewed.

Motivated by the observations in Zhao, Witten, and Shojaie (2017), we can improve the inference by refitting the rank-based model consisting of only variables selected by rank SCAD without any penalty. Under conditions that ensure variable selection consistency, that is, the true active variables are selected by rank SCAD with probability tending to 1, the refitted coefficients will be asymptotically normal and one can thus use Wald-type inference procedure. In simulations, we observe that refitting improves the coverage of confidence intervals (see Section 2.3).

One condition to ensure variable selection consistency is the beta-min condition, which states that the minimal signal strength (magnitude of nonzero coefficients) need to be larger than the estimation error. This can be stringent in the presence of weak signals. Therefore, in the next section we seek to develop inference without this condition. However, we note that the variable selection consistency requirement can be relaxed, as long as a fixed set of variables is selected with probability approaching 1 (Zhao, Witten, and Shojaie 2017).

2.2. De-correlated score and one-step estimator

The de-correlated score approach (Ning and Liu 2017) offers a general framework for statistical inference with regularized estimators in the high-dimensional setting. The original proposal works with a loss function that is second-order differentiable. However, the second derivative of the loss function in (1) does not exist on certain hyperplanes, and is 0 elsewhere. It is thus not straightforward to define an information matrix using the second derivatives. We thus aim to find a surrogate for the information.

Inspired by the statistical properties of rank-based estimators in low dimensions (Hettmansperger and McKean 2010), we propose to use the covariance matrix of X, denoted by Σ. The score function for β_j is given by,

$$S_j(\beta) = -\frac{\sqrt{12}}{n}\sum_{i=1}^{n} x_{ij}\left[\frac{R(Y_i - x_i^T\beta)}{n+1} - \frac{1}{2}\right] \quad (2)$$

and the de-correlated version is

$$\tilde{S}_j(\beta) = -\frac{\sqrt{12}}{n}\sum_{i=1}^{n}\left(x_{ij} - w_j^T x_{-j}\right)\left[\frac{R(Y_i - x_i^T\beta)}{n+1} - \frac{1}{2}\right],$$
$$w_j^T = \Sigma_{j,-j}\Sigma_{-j,-j}^{-1}. \quad (3)$$

To test the hypothesis that $\beta_j = 0$, we evaluate the de-correlated score $\tilde{S}_j$ at $\hat{\beta}^0$, which can be the rank-based lasso or SCAD estimator under the reduced model. Alternatively, we propose a one-step estimator based on this de-correlated score function, which can be used for hypothesis testing as well as constructing confidence intervals. Given an initial estimator $\hat{\beta}$, which can be the rank lasso or rank SCAD, we define

$$\tilde{\beta}_j = \hat{\beta}_j - \hat{\tau}\tilde{S}(\hat{\beta})\hat{v}_j^T X^T X\hat{v}_j/n, \;\; \hat{v}_j^T = (1, -\hat{w}_j^T). \quad (4)$$

Here, $\hat{\tau}$ is an estimate of a scale parameter τ defined as $\tau^{-1} = \sqrt{12}\int f^2(u)du$, where $f(\cdot)$ is the density of the error ϵ. This scale parameter characterizes the relative efficiency of the rank-based procedure compared to ordinary lasso.

In simulations (see Table 2), we observe that the one-step estimator based on rank lasso does not perform as well compared with the one-step rank SCAD. This suggests that the convergence of rank lasso estimator is probably not fast enough and thus there might still be correlation among score functions for different β's. The difficulty of a formal derivation of the convergence rate requirement lies in the non-differentiability of the loss function. Addressing this challenge could be a fruitful direction for future research.

Finally, we note that to implement the one-step estimators in practice, both the projection w and the scale parameter τ have to be estimated. The projection can be estimated by penalized regression if the precision matrix of X is sparse. However, the estimation of τ is less straightforward as it depends on the density function of the error.

Table 2. Coverage of Wald-type confidence intervals, for the three active variables and average over inactive variables, average over 500 replications.

	$n = 100$			
	rSCAD	rSCAD (refit)	onestep rLasso	onestep rSCAD
β_1	0.80	0.92	0.76	0.92
β_2	0.83	0.93	0.49	0.91
β_3	0.84	0.94	0.72	0.94
Inactive	–	–	0.93	0.93
	$n = 200$			
	rSCAD	rSCAD (refit)	onestep rLasso	onestep rSCAD
β_1	0.94	0.94	0.82	0.94
β_2	0.93	0.93	0.56	0.94
β_3	0.96	0.96	0.84	0.96
Inactive	–	–	0.94	0.94

2.3. Simulations

We take $n \in \{100, 200\}, p = 400$, and generate X from a mean-zero multivariate normal distribution with covariance matrix Σ, with $\Sigma_{ij} = 0.5^{|i-j|}$. We generate the outcome Y as $Y = X^T\beta_0 + \epsilon$, with $\beta_0 = (\sqrt{3}, \sqrt{3}, \sqrt{3}, 0, \ldots, 0)^T$ and $\epsilon \sim \sqrt{2}t_4$. For all the proposed inference procedure, we use the true value for the scale parameter τ. We use nodewise regression with lasso to estimate the projection w for the one-step estimators.

Table 2 shows the coverage of Wald-type confidence intervals over 500 replications for the 3 active variables and the inactive ones, based on the rank SCAD and refitted rank SCAD (selected variables only) as well as one-step rank lasso and one-step rank SCAD. We observe that refitting generally improves the inference; and for the one-step estimators, using rank SCAD as the initial estimator leads to better performance than using rank lasso.

In additional simulation experiments (not included for brevity) we investigated how the performance of one-step rank lasso changes (i) with larger sample sizes; (ii) with less correlations among covariates; or (iii) when the true covariance matrix is used for projection. Among these, only larger sample sizes result in improved coverage, which further suggests that the rate of convergence of rank lasso may not be fast enough for inference. Investigating these issues and developing efficient inference procedures for rank-based estimators would be an important direction of future research.

Funding

This work was partially supported by grants R01HL141989 and R01GM133848 from the National Institutes of Health.

References

Chen, S., Witten, D. M., and Shojaie, A. (2015), "Selection and Estimation for Mixed Graphical Models," *Biometrika*, 102, 47–64. [1717]

Drton, M., and Maathuis, M. H. (2017), "Structure Learning in Graphical Modeling," *Annual Review of Statistics and Its Application*, 4, 365–393. [1717]

Fan, J., Liu, H., Ning, Y., and Zou, H. (2017), "High Dimensional Semiparametric Latent Graphical Model for Mixed Data," *Journal of the Royal Statistical Society*, Series B, 79, 405–421. [1717]

Friedman, J., Hastie, T., and Tibshirani, R. (2008), "Sparse Inverse Covariance Estimation With the Graphical Lasso," *Biostatistics*, 9, 432–441. [1717]

Hettmansperger, T. P., and McKean, J. W. (2010), *Robust Nonparametric Statistical Methods*, Boca Raton, FL: CRC Press. [1718]

Lederer, J., and Muller, C. L. (2015), "Don't Fall for Tuning Parameters: Tuning-Free Variable Selection in High Dimensions With the TREX," in *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, pp. 2729–2735. [1717]

Lin, L., Drton, M., and Shojaie, A. (2016), "Estimation of High-Dimensional Graphical Models Using Regularized Score Matching," *Electronic Journal of Statistics*, 10, 806. [1717]

Liu, H., Han, F., Yuan, M., Lafferty, J., and Wasserman, L. (2012), "High-Dimensional Semiparametric Gaussian Copula Graphical Models," *The Annals of Statistics*, 40, 2293–2326. [1717]

Liu, H., Lafferty, J., and Wasserman, L. (2009), "The Nonparanormal: Semiparametric Estimation of High Dimensional Undirected Graphs," *Journal of Machine Learning Research*, 10, 2295–2328. [1717]

Maathuis, M., Drton, M., Lauritzen, S., and Wainwright, M. (2018), *Handbook of Graphical Models*, Boca Raton, FL: CRC Press. [1717]

Meinshausen, N., and Bühlmann, P. (2006), "High-Dimensional Graphs and Variable Selection With the Lasso," *The Annals of Statistics*, 34, 1436–1462. [1717]

Ning, Y., and Liu, H. (2017), "A General Theory of Hypothesis Tests and Confidence Regions for Sparse High Dimensional Models," *The Annals of Statistics*, 45, 158–195. [1718]

Sabourin, J. A., Valdar, W., and Nobel, A. B. (2015), "A Permutation Approach for Selecting the Penalty Parameter in Penalized Model Selection," *Biometrics*, 71, 1185–1194. [1717]

Sun, T., and Zhang, C.-H. (2012), "Scaled Sparse Linear Regression," *Biometrika*, 99, 879–898. [1717]

Voorman, A., Shojaie, A., and Witten, D. (2014), "Graph Estimation With Joint Additive Models," *Biometrika*, 101, 85–101. [1717]

Wang, L., Peng, B., Bradic, J., Li, R., and Wu, Y. (2020), "A Tuning-Free Robust and Efficient Approach to High-Dimensional Regression," *Journal of the American Statistical Association* (to appear). [1717]

Xue, L., and Zou, H. (2012), "Regularized Rank-Based Estimation of High-Dimensional Nonparanormal Graphical Models," *The Annals of Statistics*, 40, 2541–2571. [1717]

Yu, S., Drton, M., and Shojaie, A. (2019), "Generalized Score Matching for Non-Negative Data," *Journal of Machine Learning Research*, 20, 1–70. [1717]

Yuan, M., and Lin, Y. (2007), "Model Selection and Estimation in the Gaussian Graphical Model," *Biometrika*, 94, 19–35. [1717]

Zhao, S., Witten, D., and Shojaie, A. (2017), "In Defense of the Indefensible: A Very Naive Approach to High-Dimensional Inference," to appear in *Statistical Science*. [1718]